УДК 621.326
Мигаль З. – ст. гр. СП-42
Тернопільський національний технічний університет імені Івана Пулюя

РОЗРОБКА МОДЕЛІ МАШИННОГО НАВЧАННЯ ДЛЯ ГЕНЕРАЦІЇ ТЕКСТОВИХ ОПИСІВ ЗОБРАЖЕНЬ НА ОСНОВІ VISION-LANGUAGE ПІДХОДІВ

Науковий керівник: к.ф.-м.н., доцент Цебрій О. Р.

Myhal Z.
Ternopil Ivan Puluj National Technical University

DEVELOPMENT OF A MACHINE LEARNING MODEL FOR GENERATING IMAGE CAPTIONS BASED ON VISION-LANGUAGE APPROACHES

Supervisor: PhD in Physics and Mathematics, Associate Professor Tsebrii O.

Ключові слова: генерація тексту, комп'ютерний зір, vision-language
Keywords: text generation, computer vision, vision-language

Задача автоматичної генерації текстових описів зображень є важливим напрямом сучасних досліджень у сфері штучного інтелекту, оскільки вона поєднує методи комп'ютерного зору та обробки природної мови. Такі системи знаходять застосування у допоміжних технологіях для людей з порушеннями зору, системах пошуку зображень, автоматичному тегуванні контенту та цифрових асистентах. Основною складністю задачі є необхідність коректного поєднання візуальної інформації з мовними конструкціями, що потребує глибокого розуміння як зображення, так і контексту.

Сучасні підходи базуються на використанні глибоких нейронних мереж, зокрема згорткових нейронних мереж для витягування ознак із зображень та трансформерних моделей для генерації тексту. Архітектури типу encoder-decoder дозволяють перетворювати візуальні ознаки у послідовності слів, формуючи осмислені описи. Використання attention-механізмів забезпечує можливість фокусування на окремих ділянках зображення під час генерації кожного слова, що підвищує якість та релевантність описів. Окрім класичних підходів, активно розвиваються vision-language моделі, такі як CLIP, BLIP та їх похідні, які навчаються на великих мультимодальних датасетах і здатні ефективно узгоджувати текстові та візуальні представлення. Такі моделі демонструють високу узагальнювальну здатність і можуть застосовуватись у широкому спектрі задач без значного донавчання.

Важливим етапом є підготовка даних, яка включає формування пар «зображення–опис», очищення тексту та нормалізацію зображень. Якість датасету безпосередньо впливає на результати моделі, оскільки некоректні або неповні описи можуть призводити до помилок у генерації. Для оцінювання якості використовуються метрики, такі як BLEU, METEOR, CIDEr та ROUGE, що дозволяє комплексно оцінити відповідність згенерованого тексту реальному опису [1].

Важливим є врахування контексту та семантичної узгодженості описів. Модель повинна не лише ідентифікувати об'єкти на зображенні, але й правильно описувати їх взаємодію, просторове розташування та дії. Для цього використовуються механізми

попереднього навчання на великих корпусах даних, що дозволяє покращити мовну складову системи.

Додатково важливим аспектом є вибір способу представлення візуальних ознак, які передаються у мовну модель. Сучасні підходи використовують як глобальні представлення зображення, так і локальні ознаки окремих регіонів, що дозволяє більш точно описувати складні сцени. Використання регіональних фіч або patch-представлень у трансформерних моделях забезпечує кращу деталізацію описів та підвищує здатність моделі враховувати дрібні об'єкти та їх взаємозв'язки.

Особливу роль відіграє процес узгодження візуального та текстового простору ознак. Для цього застосовуються контрастивні методи навчання, які дозволяють зблизити відповідні пари «зображення–текст» та віддалити невідповідні. Такий підхід підвищує якість семантичного розуміння та дозволяє моделі краще узагальнювати нові дані. Крім того, це відкриває можливості для використання моделі у суміжних задачах, таких як пошук зображень за текстовим запитом або навпаки.

Також важливим є врахування якості та різноманітності текстових описів, оскільки одна сцена може мати декілька коректних варіантів опису. Для вирішення цієї проблеми застосовуються методи генерації з використанням стохастичних підходів, таких як beam search або sampling, що дозволяє отримувати більш природні та варіативні результати. Це підвищує гнучкість системи та робить її більш придатною для реальних застосувань, де важлива не лише точність, але й природність сформованого тексту.

Окрему увагу приділяють оптимізації моделей та їх впровадженню у практичні системи. Методи стиснення моделей, такі як квантизація та дистиляція знань, дозволяють зменшити обчислювальні витрати та забезпечити роботу в реальному часі. Це відкриває можливості використання таких систем у мобільних додатках та веб-сервісах [2].

Таким чином, поєднання методів комп'ютерного зору та обробки природної мови у рамках vision-language підходів формує цілісну технологічну основу для побудови ефективних систем генерації текстових описів зображень. Важливим є не лише використання окремих моделей, а їх узгоджена інтеграція у єдину архітектуру, де візуальні ознаки коректно трансформуються у мовні представлення з урахуванням контексту та семантики. Завдяки використанню попередньо навчених мультимодальних моделей та великих датасетів досягається висока якість узагальнення, що дозволяє системам працювати з різними типами зображень та сценаріями застосування без значного донавчання. У результаті формується інтелектуальна система, яка може бути інтегрована у веб-сервіси, мобільні додатки та інші цифрові платформи, забезпечуючи більш природну та ефективну взаємодію людини з інформаційними системами. Це сприяє розширенню можливостей доступу до інформації, покращенню якості цифрових сервісів та створенню нових підходів до обробки й інтерпретації мультимедійних даних.

1. Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. — ICML.
2. Li, J., Li, D., Savarese, S., Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training. — ICML.
3. Alayrac, J.-B., Donahue, J., Luc, P., et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. — NeurIPS.